DOCUMENT RESUME

ED 170 332 TM 008 558

AUTHOR Arter, Judith A.; Estes, Gary D.
TITLE A Model for Developing Local Norms with a Standardized Achievement Measure for Use with Local Program Evaluation: Procedures and Effects.
PUB DATE Mar 78
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Norm Referenced Tests; *Norms; Program Evaluation; Research Methodology; Senior High Schools; Standardized Tests; *Student Ability; Student Testing; *Test Validity; Time Factors (Learning)
IDENTIFIERS *Comprehensive Tests of Basic Skills; *Local Norms; Phoenix Union High School System AZ

ABSTRACT

The validity of the Comprehensive Tests of Basic Skills (CTBS) and problems associated with inappropriate norms were investigated in eleven senior high schools in Phoenix, Arizona. CTBS validity was assessed in three ways: (1) correlation between test scores and number of courses taken in that subject; (2) correlation between CTBS scores and locally-developed criterion referenced tests; and (3) relationship between course difficulty and CTBS scores. A reverse norming procedure was used to obtain a constant normed comparison group--a group which did not increase inability in successive grades. This resulted in the analysis of data from 631 ninth and tenth graders in two Title I schools. Results of the validity studies indicated that correlations between CTBS scores and courses taken were high for science and mathematics courses, lower for English courses, and negative for reading courses. CTBS correlations with locally developed tests ranged from .38 to .83. Students in more difficult science and mathematics courses had higher test performance. Regarding the use of reverse norming, no clear pattern of differences emerged between restricted and nonrestricted norms. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED170332

A Model for Developing Local Norms with a Standardized Achievement Measure for Use with Local Program Evaluation:

Procedures & Effects

Judith A. Arter
Phoenix Union High School System

Gary D. Estes
Northwest Regional Education Laboratory

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Judith A. Arter

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Printed in U. S. A.

TM008 558

*Paper presented at the American Education Research Association National Convention, Toronto, March 27-31, 1978.

A MODEL FOR DEVELOPING LOCAL NORMS WITH A STANDARIZED ACHIEVEMENT MEASURE FOR USE WITH LOCAL PROGRAMS EVALUATION PROCEDURES AND EFFECTS

JUDITH A. ARTER GARY G. ESTES

Local program evaluation is often based on achievement testing employing standardized tests. Interpretation of the results of such evaluations (e.g. the norm referenced designs used in Title I evaluations) are often made difficult because of several factors. First, standardized tests may not adequately test local curriculum. Since local emphasis may be different from that measured by a test, true gains made in local programs may not show up on a standardized test. Consequently an incorrect judgment may be made that a program is unsuccessful.

A second problem associated with the use of standardized tests in local program evaluation relates to the norming samples of such tests. Although standardized tests use national samples of students for norming purposes, local situations may produce a student population that does not correspond closely to the test norming sample (Wood & Tallmadge, 1976). Thus, for example, if a local school population is, in general, below average, year gains as measured by a test might be less than those in the norming sample, even under the best of conditions. Thus, again there might be an incorrect conclusion that a program is ineffective (Fagan & Horst, 1976).

Third, testing in the schools often takes place at times other than the dates on which the norms of a standardized measure are based. Norms often are made by testing at one time of the year and then interpolating between these fixed points in order to get norms for other times of the year. This assumes a fixed linear learning growth with no summer forgetting (Burrill, undated). This could result in spuriously inflating or deflating progress.

A fourth problem relates again to summer forgetting (Estes & Faddis, 1977). It is possible, if doing fall to spring evaluation, to have overall gains each year, but show no general increase in knowledge over many years. This situation would occur if students forgot everything over the summer that they had learned during the previous school year.

* Paper presented at the AERA national convention, Toronto, March 27-31, 1978.

Finally, a program may incorrectly show low gains over time just because the norming groups to which the students are being compared are increasingly improving. This is a consideration especially when dealing with extreme or unusual groups, such as Title I or compensatory students, which typically are not included in national norms, but which consistently are given standardized tests for purposes of program evaluation. If the progress of such students is compared to national norms, there may be a spurious decrease in performance because of the dropping out through time of the lower ability students included in the national norms each year. This may be true particularly at the high school level. This results in a non-constant comparison group in which progressively higher ability students constitute the norming group.

The Phoenix Union High School System (PUHSS), which uses the Comprehensive Test of Basic Skills (CTBS) as a measure of academic growth in evaluating its special programs, has initiated procedural and data management changes, and validity studies in order to address the above threats to clearer interpretation of evaluation results.

For example, the Phoenix Union High School System has instituted two procedural changes designed to clarify evaluation results. The first change is provision for yearly, October-to-October testing. This ensures that judgments concerning program efficacy are not based on fall-to-spring gains which may disappear over the summer.

The second change was the development of local norms. These norms are broken down by grade level and ethnic group. They provide for conversion from raw scores, standard scores or stanines to percentiles. Since the CTBS is normed for January-to-January or April-to-April testing (Tallmadge and Wood, 1976), local norms based on October testing are more accurate for measuring change in the PUHSS district. Local norms also provide a comparison group that more closely resembles the groups being evaluated. In addition, norms for ethnic groups are useful for program evaluation and to determine whether there is ethnic test bias on the instrument.

This paper will examine two other issues which are threats to clear interpretation of evaluation results: validity of the CTBS for local use, and the problem of progressively better ability students in the norming

sample. The information presented on these topics is meant to be suggestive of the type of studies that school districts might carry out in order to examine the usefulness of a test in local evaluation.

Method

School Population

The PUHSS is a large metropolitan school district which serves approximately 30,000 ninth through twelfth grade students in eleven high schools. The System serves about 1,500 Title I students, 300 Bilingual students and 2,200 special education students. There are about 20,000 white students, 3,000 black students, 6,000 Spanish surname students, and 1,000 miscellaneous other students. The district encompasses middle and low socio-economic statuses.

Test Validity

Three analyses were undertaken which addressed the issue of the validity of the CTBS for use in the PUHSS. These were, a) correlations between CTBS subtest scores and the number of district courses taken in that subject area; b) correlations between scores made on district developed criterion referenced tests (Minimal Reading Proficiency Assessment, MRPA; and Minimal Math Proficiency Assessment, MMPT*) and CTBS subtest scores; and c) the relationship between course difficulty (many district courses have sections for slow, regular and advanced students) and test scores. These are meant to be suggestive of the types of studies which may be undertaken by other districts.

* Tests developed by, and data provided courtesy of Dr. Coleen R. Goodwin, Acting Director of Research and Planning, PUHSS.

The first and third analyses, i.e., correlation between CTBS scores and number of district courses taken in various subject areas, and comparison of CTBS scores for students in the various difficulty levels of the same class, used as subjects all seniors (N = 4849) who took the CTBS in October, 1976. CTBS (Level 4) expanded standard scores were used in these analyses. Number of courses taken and their difficulty level was collected by student self report and from the district master file at the time the test was given.

Analysis two, correlations between district criterion-referenced tests and CTBS scores were based on a sample of freshmen at each school (N = 658). Both the CTBS and district criterion-referenced tests were given in October, 1976. CTBS stanines were used in the computations.*

Reverse Norming

In order to examine the possibility of spurious low program gains when comparing progress to a national norming sample that may be increasing in ability in succeeding grades, a procedure which we labeled "reverse norming" was used. Reverse norming consists of identifying all students for whom data is available for both points in time during which growth is to be measured. For example, in order to measure longitudinal growth of Title I students from the ninth to the twelth grade, ninth and twelth grade district norms would be developed using only those students with test results at both points in time (restricted district norms). The Title I students with ninth and twelth grade test results would be compared to these restricted norms. The procedure results in a normed comparison group that is constant.

* Tests developed by, and data provided courtesy of Dr. Coleen R. Goodwin, Acting Director of Research and Planning, PUHSS.

Because the PUHSS has only initiated district-wide CTBS testing for the last two years, the results of "reverse norming" are reported for the sophomore classes at two Title I schools projected back on the freshmen classes at those schools (School 1, N = 237, School 2, N = 394). Four sets of local norms were generated for each school: 9th and 10th grade unrestricted (i.e., the entire class) norms, and 9th and 10th grade restricted (i.e., that subpopulation of the schools constant in both the 9th and 10th grades) norms.

If dropouts are not random, then one would expect that 10th grade restricted and nonrestricted standard score to percentile norms would be approximately the same, but that 9th grade restricted norms would show higher standard scores corresponding to the same percentile ranking of 9th grade unrestricted norms. This result would occur because 10th grade restricted and unrestricted samples should be equivalent in achievement, but the 9th grade restricted sample should be higher in ability than the 9th grade unrestricted sample because of the dropping out of lower ability students.

It would also be expected that the gains for a group of students might be different when using restricted and unrestricted norms.

Results and Discussion

Test Validity

Correlation Between CTBS Subtest Scores and Number of District Courses in Each Area. These correlations are shown in Table 1. All correlations are significant at an alpha level of .001, except where noted by double or triple asterisks. Intercorrelations between CTBS subtests are high, as expected. The relationship between numbers of science and math classes taken is also strong. The relationships between number of science and math classes on the one hand and number of English and Reading classes on the other, are very low. This is presumably because only students who need remedial help in reading (and to a lesser extent English) take more than one reading class. And, students requiring remedial reading and English would not be expected to take many science and math classes.

Table 1

Correlations Between CTBS Subtest Standard Scores
and Number of Courses Taken in Each Area
All Seniors, 1976 (N=4849)*

		Number of Classes						CTBS Standard Score				
		Math	Science	English Semester	English Trimester	Reading Semester	Reading Trimester	Math	Science	English	Reading	Composite
No. Classes:	Math											
	Science	.499 (N=4849)										
	English (Sem)	.120 (N=3331)	.125 (N=3331)									
	English (Tri)	.065 (N=1429)	.088 (N=1429)	.504 (N= 183)								
	Reading (Sem)	-.022 (N=2285)	.002*** (N=2285)	.178 (N=2244)	.247** (N= 132)							
	Reading (Tri)	-.076 (N= 868)	.058** (N= 868)	.126 (N= 141)	.133 (N= 843)	.205** (N= 104)						
CTBS:	Math	.478 (N=3916)	.296 (N=3916)	.087 (N=2733)	.089 (N=1179)	-.238 (N=1884)	-.290 (N= 701)					
	Science	.322 (N=4380)	.265 (N=4380)	.061 (N=3008)	.093 (N=1329)	-.249 (N=2073)	-.267 (N= 808)	.745 (N=3776)				
	English	.279 (N=4312)	.184 (N=4312)	.058 (N=3015)	.151 (N=1293)	-.254 (N=2076)	-.222 (N= 796)	.720 (N=3771)	.680 (N=4006)			
	Reading	.290 (N=4399)	.213 (N=4399)	.078 (N=3068)	.113 (N=1323)	-.251 (N=2098)	-.243 (N= 807)	.716 (N=3765)	.758 (N=4062)	.829 (N=4139)		
	Composite	.404 (N=3665)	.264 (N=3665)	.071 (N=2565)	.102 (N=1122)	-.262 (N=1781)	-.296 (N= 667)	.912 (N=3665)	.799 (N=3599)	.907 (N=3665)	.910 (N=3664)	

* All $p < .001$, except where marked with a double or triple asterisk
** $p < .05$
***nonsignificant

Correlations between number of science and math classes taken and CTBS scores on all subtests are quite high (between r = .184 and r = .498). This may indicate that knowledge obtained in district classes is cummulative and directly applicable to the CTBS, or it may mean that stronger students take more science and math. Correlations between CTBS subtests and number of English classes are much lower (r = .058 to r = .151), and correlations with number of Reading classes are negative (r = -.290 to r = -.222). The reason for the English result may be a combination of accumulating knowledge in district classes and having both slow and advanced students in English classes. The Reading result occured probably because only slow students take more than one reading class.

Correlations Between District Criterion-Referenced Tests and CTBS Scores. Table 2 shows the correlations between the MRPA, MMPT and CTBS subtest scores on those subtests most closely resembling in intent the district developed criterion-referenced tests. Not only are all of these correlations significant at an alpha level of .01, but all are well above the r = .35 usefulness cutoff recommended by Guilford (1965) and Garrett (1954).

Comparison of CTBS Scores for Students in Varying Difficulty Levels of the Same Course. Only math and science offered courses at various difficulty levels for slow, regular, and above average students. Scores for students in these classes are presented in Tables 3a and 3b. T-tests were conducted on the CTBS stanine scores in math or science for students in the varying difficulty levels of these classes. One way analyses of variance were calculated for the two science classes that had all three difficulty levels of classes. Because so many analyses were done, an alpha level of .01 should be used to test significance. For math classes, students in the harder level of a class always outperformed students in the easier level on math CTBS scores. Only one situation (Introduction to Analysis, regular vs. accelerated students) failed to reach significance ($p < .025$). It is also interesting to note the progression of means in the different classes. Means for students in more advanced classes (i.e. Algebra 3-4) are higher than means for students in more elementary classes (i.e. General Math 1-2).

Table 2

Correlations Between Selected
CTBS Subtest Stanines and District Criterion
Referenced Tests*

School	N	CTBS Reading Comprehension Stanines & MPRA	CTBS Math Computation MMPT
1	49	.38	.67
2	48	.51	.72
3	27	.59	.78
4	21	.72	.64
5	57	.53	.78
6	66	.46	.73
7	46	.69	.75
8	49	.62	.83
9	141	.56	.59
10	108	.55	.61
11	47	.71	.82

*Courtsey of Dr. Coleen R. Goodwin, Acting Director, Research & Planning, Phoenix Union High School System. All correlations were significant at $p < .01$

Table 3a

CTBS Stanines for Students in
Varying Difficulty Levels of the Same Course

Math

All Seniors, 1976 (N=4849)*

Class / CTBS Subtest/ Level		Math Computation			Math Concepts			Math Application		
		Slow	Regular	Accelerated	Slow	Regular	Accelerated	Slow	Regular	Accelerated
Gen Math 1-2	N	88	1770		93	1768**		80	1601**	
	$\bar{X}$	2.89	3.62		3.04	3.53		3.39	3.83	
	SD	1.35	1.47		1.28	1.42		1.26	1.46	
Algebra 1-2	N		1775	262		1764	255		1662	247
	$\bar{X}$		5.60	7.00		5.59	6.79		5.60	7.06
	SD		1.58	1.85		1.68	1.83		1.61	1.77
Geometry 1-2	N		1233	101		1216	93		1167	92
	$\bar{X}$		6.14	7.37		6.28	7.32		6.13	7.07
	SD		1.48	1.52		1.46	1.53		1.53	1.68
Algebra 3-4	N		426	136		419	134		406	133
	$\bar{X}$		6.68	7.63		6.72	7.54		6.42	7.70
	SD		1.50	1.37		1.61	1.28		1.63	1.38
Intro Analysis 1-2	N		16	22+		15	22**		15	22++
	$\bar{X}$		6.50	8.00		6.53	7.95		6.53	7.91
	SD		2.07	1.60		1.81	1.43		2.03	1.63

*Courtsey of Dr. Coleen R. Goodwin, Acting Director, Research & Planning, Phoenix Union High School System
All $p<.001$, unless otherwise noted.
**$p<.005$
+$p<.01$
++$p<.03$

Table 3b

CTBS Scores for Students in
Varying Difficulty Levels of the Same Course

Science

All Seniors, 1976 (N=4849)*

Course	Slow			Regular			Accelerated		
	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD
Earth Science+	15	4.13	1.30	919	5.37	1.87			
Bio 1-2**	69	3.87	1.71	3230	5.18	1.82	246	6.06	2.00
Bio 3-4 +				291	4.46	1.75	45	5.11	1.92
Chem 1-2**	40	5.70	1.44	787	6.34	1.66	199	6.86	1.66
Chem 3-4 +++				20	5.00	1.62	30	6.83	1.86
Physics 1-2 +++	2	4.50	*** .71	155	6.70	1.83	74	7.47	1.48
Physiology ++				44	5.59	1.74	6	7.10	1.83

*Courtesy of Dr. Coleen R. Goodwin, Acting Director, Research & Planning, Phoenix Union High School System

**Analysis not conducted

***These students were not included in the analysis because of the low N.

+$p < .03$

++$p < .05$

+++$p < .005$

For science classes similar patterns emerge. Students in the harder levels of a course always outperformed students in the easier levels on CTBS science score. For the five t-tests and two ANOVAS computed, four (Biology 1-2, Chemistry 1-2, Chemistry 3-4 and Physics 1-2) were significant at $p<.01$. The others were significant at $p<.05$. In science, the progression of means from the more elementary classes to the more advanced is not as apparent as in the math classes.

Reverse Norming

Tables 4a (School #1) and 4b (School #2) compare the CTBS composite standard scores at several percentiles for ninth and tenth grade restricted and unrestricted norms. At school #2 but not school #1 the prediction was demonstrated that tenth grade restricted and unrestricted populations are approximately the same in distribution of standard scores.

To test the hypothesis that using restricted and unrestricted norms to measure progress would yield different measures of progress, we have compared the amount of standard score increase required to maintain the same percentile between 9th and 10th grade restricted norms, and 9th and 10th grade unrestricted norms. These changes appear in the last two columns of Tables 4a and 4b.

For example, at school #2 using unrestricted norms, an increase of 21 standard score points are needed for a group who began at the 10th percentile to remain at the 10th percentile. For restricted norms, 31 points are required. However, by examining these differences at various percentiles and between the two schools, it is clear that no pattern of differences emerges.

Table 4a

CTBS Composite Standard Scores
at Various Percentiles

High School #1

Percentile	9th Grade Unrestriced	9th Grade Restricted	10th Grade Unrestricted	10th Grade Restricted	Change Required to Remain at Same Percentile		Differences
					Unrestricted	Restricted	U - R
99	776	780	827	827	51	47	4
90	642	642	683	687	41	45	-4
80	590	591	628	632	38	41	-3
70	552	554	595	604	43	50	-7
60	536	541	571	582	35	41	-6
50	513	513	558	560	45	47	-2
40	492	499	538	546	46	47	-1
30	466	471	511	515	45	44	1
20	443	446	477	477	34	31	3
10	402	405	434	437	32	32	0
	N=296	N=203 (237)	N=299	N=221 (237)			

Table 4b

CTBS Composite Standard Scores
at Various Percentiles

High School #2

Percentile	9th Grade Unrestriced	9th Grade Restricted	10th Grade Unrestricted	10th Grade Restricted	Change Required to Remain at Same Percentile		Differences
					Unrestricted	Restricted	U - R
99	650	648	740	737	90	89	1
90	591	597	656	652	65	55	10
80	556	560	611	611	55	51	4
70	528	536	579	577	51	41	10
60	513	520	552	552	39	32	7
50	484	497	522	517	38	20	8
40	458	466	494	492	36	26	10
30	428	437	466	469	38	32	6
20	402	408	431	437	29	29	0
10	377	381	398	412	21	31	-10
	N=375	N=362 (394)	N=515	N=357 (394)			

Summary and Conclusions

The CTBS appears to be a valid measure of achievement in the PUHSS. Correlations between CTBS subtest scores and number of System courses taken in the various areas closely fits our knowledge of the class-taking patterns of students. Correlations between CTBS reading and math scores and System made reading and math tests are usefully high. Finally, students in higher difficulty levels of a single math or science class always outperform (on the CTBS) students in lower difficulty levels. And in math, students in more advanced classes outperform students in less advanced classes.

These analyses and results might be viewed as trivial by some; i.e., it would only be expected that stronger students in more difficult courses will score higher on almost any test. While the authors would not argue with this position, these questions were posed by curriculum consultants and teachers who doubted the validity of standardized tests in being sensitive to differences among students or levels of skills. Thus, the results and anaylses were useful in resolving some questions about the CTBS's usefullness to the District and in improving confidence is the use of the CTBS.

The results of the "reverse norming" analyses presented in this paper did not support the conclusions that the procedure will yield a more sensitive norm group against which to measure progress. It is possible that the dropping out of low level students will not influence the results over a one year span, but will make a difference over longer periods (e.g. the four years of high school). Thus, although it is possible that a changing norm group, such as that included in national samples, will not be a problem, it is recommended that other districts conduct similar analyses in order to confirm or disconfirm the results reported here.

References

Burrill, L.E.

How a standarized achievement test is built. Test Service Notebook 125, New York: Harcourt, Brace, Jovanovich, Inc.

Estes, G. & Faddis, B.J.

Fall to spring versus fall to fall evaluations in a large high school systems Title I program with a comparison group design. Paper presented at AERA, Toronto, 1978.

Fagan, B.M. & Horst, D.P.

Selecting a norm-referenced test. ESEA Title I Evaluation and Reporting System, Technical Paper No. 4, RMC Research Corp., Mountain View, California, October, 1976.

Garrett, H.E.

Statistics in psychology and education. New York: Longmans Green, 1954.

Gailford, J.P.

Fundamental statistics in psychology and education. New York: McGraw-Hill, 1956.

Tallmadge, G.K. & Wood, C.T.

Characteristics of eight commonly used, nationally normed tests. ESEA Title I Evaluation and Reporting System, Technical Report No. 5, RMC Research Corp., Mountain View, California, October 1976.

Wood, C.T. & Tallmadge, G.K.

Local norms. ESEA Title I Evaluation and Reporting System, Technical Paper No. 7, RMC Research Corp., Moutnain View, Califronia, October, 1976.

Format:

Burrill, L.E., How a standardized achievement test is built. Test Service Notebook 125, New York: Harcourt, Brace, Jovanovich, Inc.